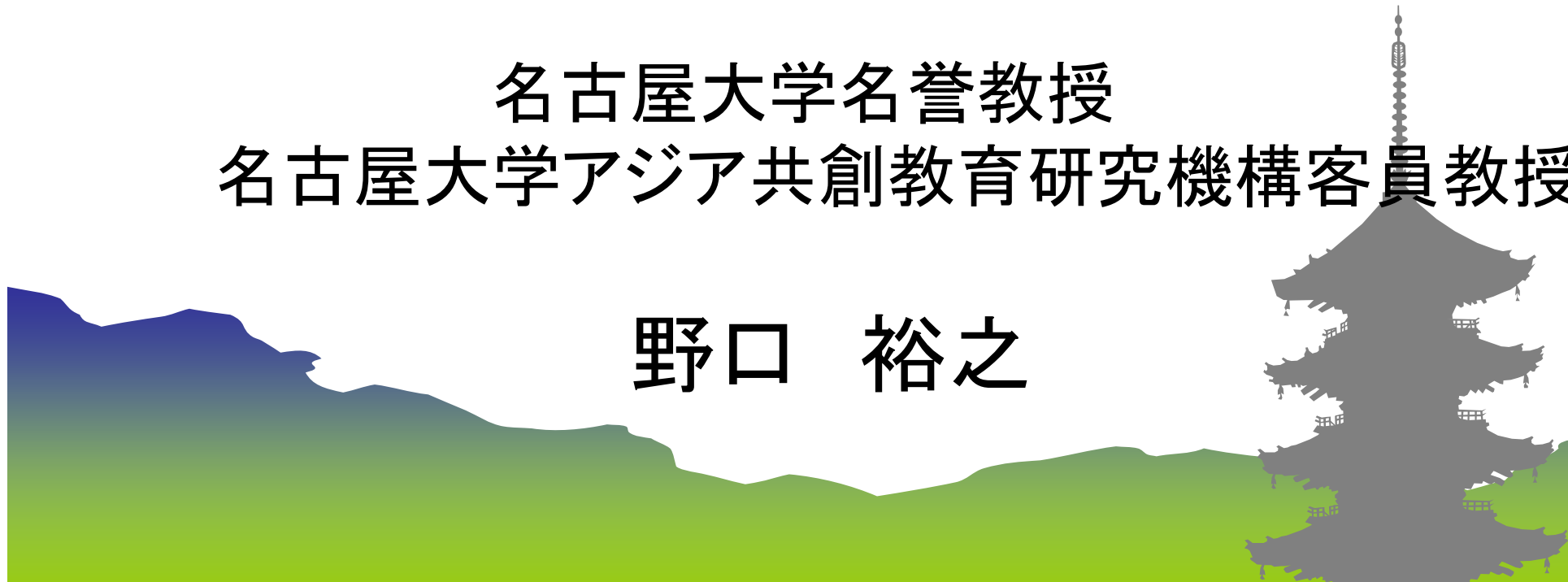


IRTとCBTの光と影 —高大接続改革の夢か現か幻か—

名古屋大学名誉教授
名古屋大学アジア共創教育研究機構客員教授

野口 裕之



本日の流れ

1. 高大接続改革とテスト
2. IRTとは何か
3. CBTとは何か
4. 英語4技能外部試験をどう位置付けるか
5. 本日のまとめ

1. 高大接続改革とテスト

大学入学共通テスト

1. 大学入試センター試験に代わるテスト:「共通テスト」と略される
2. 大学入学希望者を対象に、高等学校段階における基礎的な学習の達成の程度を判定し、大学教育を受けるために必要な能力について把握することを目的とする。

このため、各教科・科目の特質に応じ、知識・技能を十分に有しているかの評価も行ないつつ、思考力・判断力・表現力を中心に評価を行なうものとする。

3. 大学入試センターが問題の作成、採点その他一括して処理することが適当な業務等を行なう。
4. 【英語】大学入学者選抜においても、「読む」「聞く」「話す」「書く」の4技能を適切に評価するため、共通テストの枠組みにおいて、既に民間事業者等により広く実施され、一定の評価が定着している資格・検定試験を活用する。
5. 大学入試センターが必要な水準・要件を満たしているものを認定し、その試験結果および CEFR の段階別成績表示を要請のあった大学に提供する。

6. 国は、CEFRの段階別成績表示による対照表を提示する。
7. 各大学は、認定試験の活用や、個別試験により英語4技能を総合的に評価するように努める。
8. CBTの導入については、引き続きセンターにおいて、導入に向けた調査・検討を行なう。

「大学入学共通テスト」実施方針(案) (2017.05.16.文科省公表)より抜粋
文言は適宜変更した。

- 少なくとも、この実施方針(案)に、
 - i) IRTの3文字は出て来ない
 - ii) CBTの3文字は一か所だけ
 - iii) 英語の外部資格・検定試験を通して、IRTとCBTとが関係しているだけになっている。

2. IRTとは何か

項目応答理論

(Item Response Theory : IRT)

- テスト項目や受験者集団に依存せずに、テストの受験者の能力値を算出するためのテスト理論のこと。項目反応理論とも呼ばれる。
- 古典的テスト理論 (Classical Test Theory, CTT) に対して現代テスト理論と位置づけられている。

- 古典的テスト理論では、項目の困難度を表わすのに正答率(通過率)が用いられるが、この指標は全受験者に対する正答者数の比率で定義されるため、受験者集団が異なると同一項目でもその値が異なってしまう。
- また、基本的に受験者の能力を表わすのに正答数得点が用いられるが、正答数得点は当該受験者の正答した項目数で定義されるため、同一受験者でも項目の異なるテストでは得点が異なってしまう。困難度の高い項目を多く含んだテストでは、困難度の低い項目を多く含んだテストよりも、得点が低くなる。

- これに対して項目応答理論では、
 - 1) 受験者集団に依存しないで項目の特性を、
 - 2) 解答した項目に依存しないで受験者の能力(特性)を表わすことができる。

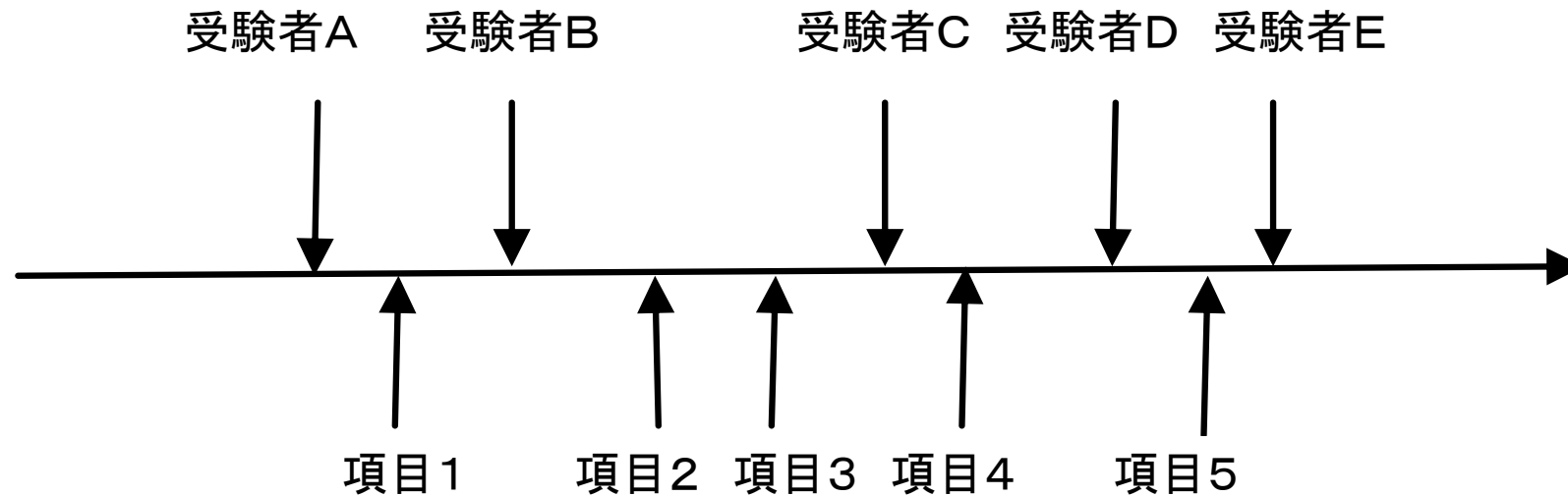
- 項目応答理論は、Lord(1952)で既にその基礎は確立され、Lord & Novick(1968)で数理的に体系化されており、TOEFLのように、ETS(Educational Testing Service)で開発実施される試験を中心に実用水準で既に用いられていた。
- 日本で注目され、公的試験などで用いられるようになったのは最近のことである。
- 米国のみならず、欧州、豪州などでも言語テストで広く用いられている。

ラッシュ・モデル

- 項目応答理論と数理的には同じモデル群に含まれるが、別の文脈から導かれたモデル。
本講演ではラッシュ・モデルも合わせて、IRTモデルと呼ぶ。
- デンマークの数学者、Rasch, G. が受験者の能力を表わす値（パラメタ）と項目の困難度を表わす値（パラメタ）を分離して独立に表現するモデルを提案した(Rasch, 1960)。
- その後、シカゴ大学の Wright, B. を中心に研究および普及活動が進められ、欧州や豪州では言語テストの分析モデルとして標準的なモデルとなっている。

IRTモデルの基本的な考え方

IRTモデルでは、受験者の能力水準と項目の困難度水準とが同一の尺度上に位置づけられて表現される



能力水準の高低に応じて5名の受験者が、また困難度水準の高低に応じて5つの項目が順に直線上に配置されている。
直線の右方向にある受験者ほど能力が高く、右方向にある項目ほど難しいことを表わしている。

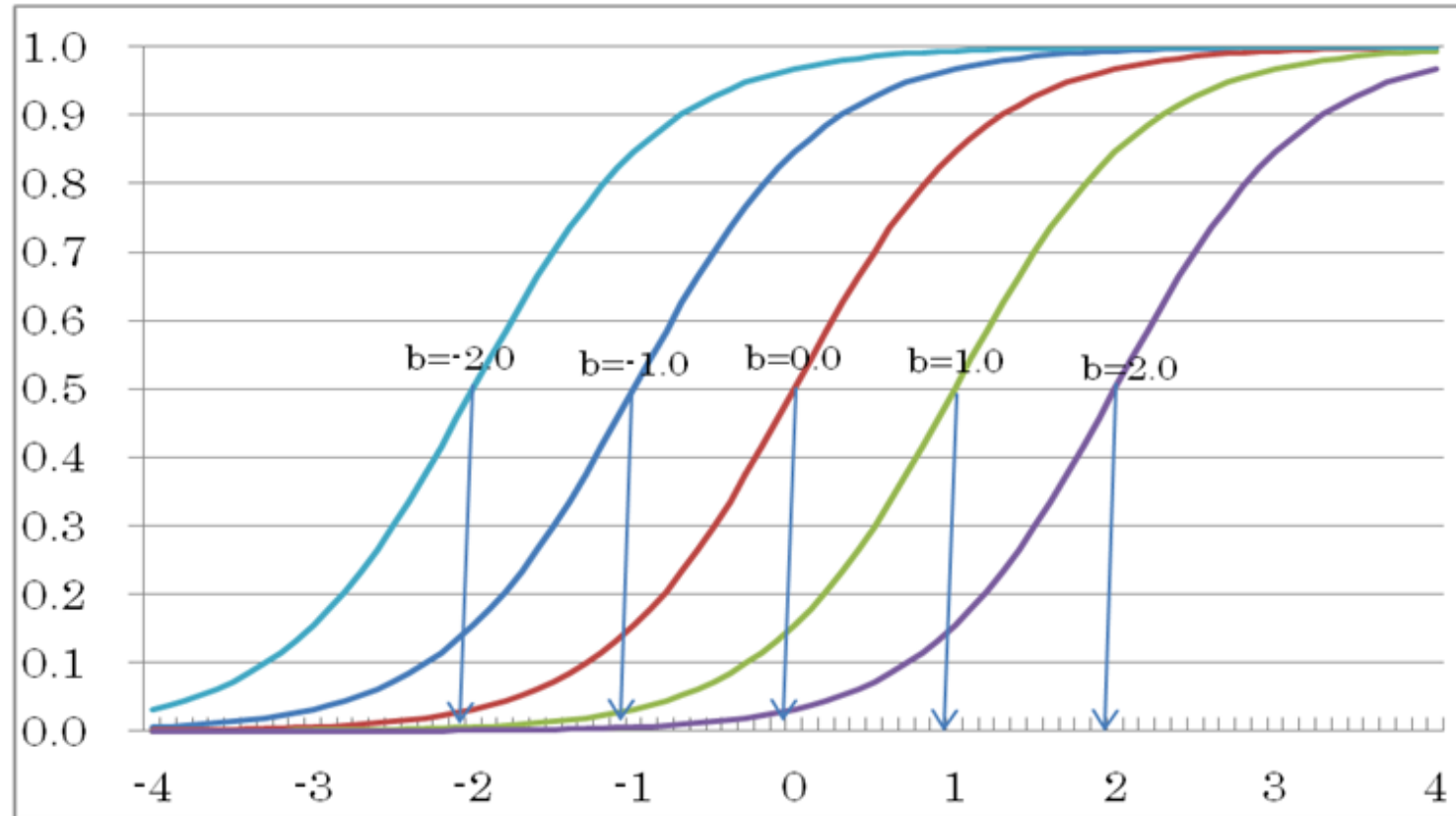
- 直線上の位置で、
受験者一項目の距離が0.0より大きいほど、
当該受験者がその項目に容易に正答でき、
0.0より小さいほどその項目に正答することが困難である。
- 受験者Eは項目2に容易に正答できるが、受験者Aが項目2に正答するのは難しい。

- 実際には、

正答のしやすさを「正答確率」で表現し、
横軸の値（潜在特性尺度値）と正答確率との関係を
表わすを表わす関数を具体的に定義する。

この関数を「項目特性関数」と呼び、実際に図示したものを「項目特性曲線」と呼ぶ。

例えば、ラッシュ・モデルの場合の項目特性曲線を、次の
スライドに示した。



横軸が図1の直線に相当し、項目の困難度および受験者の能力尺度値を表わす座標軸
縦軸が確率を示す座標軸で、項目特性曲線により横軸の値に対応する正答確率を読み取ることができる。
項目困難度 $b_j = -2, -1, 0, +1, +2$ の場合の5項目を例示した。

IRTモデル

【モデル】

項目応答理論では、項目の特性が項目特性曲線を用いて表わされる。

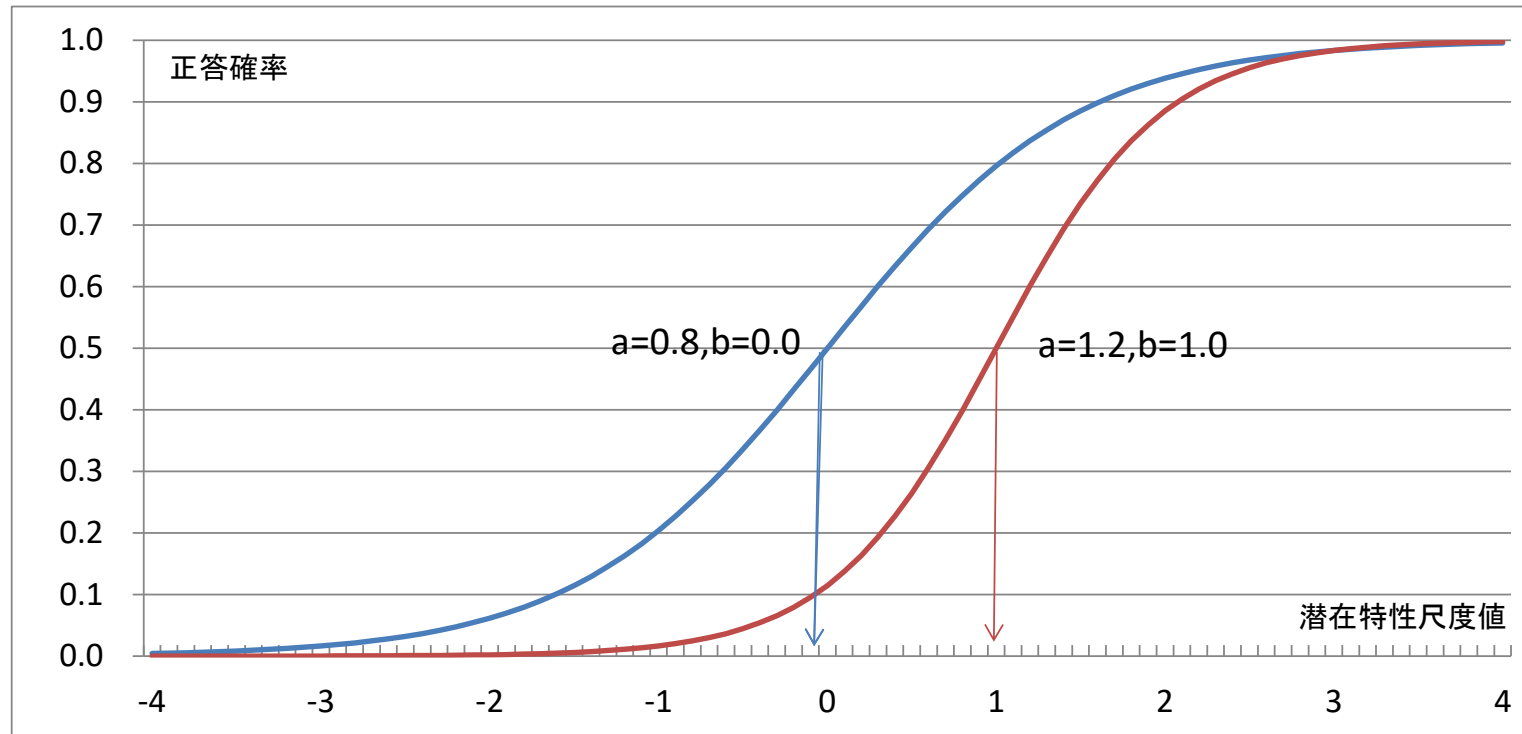
これは、項目毎に特性尺度値を横軸にとり、その特性尺度値を持つ受験者が正答する確率を縦軸方向にプロットして得られる曲線

である。

➡ 次のスライド

項目特性関数にどのような関数を用いるかに応じて、具体的なモデルが決まる。

項目特性曲線(の例)



IRTモデルの代表的なもの

- 「2値型応答モデル」

受験者の応答が正答か誤答など2つの段階で表わされる場合に適用される。

- 「多値型応答モデル」

受験者の応答に部分得点が与えられるなど段階づけられたカテゴリで表わされる場合に適用される。

- よく用いられる「2値型応答モデル」の代表的なものに、
2パラメタ・ロジスティック・モデルがある。

2パラメタ・ロジスティック・モデル

このモデルでは、項目特性曲線が

$$P_j(\theta) = \frac{1}{1 + \exp\{-1.7a_j(\theta - b_j)\}}$$

という関数で表わされる。 ➡ 次のスライド

ここで、**添字のjは項目番号**を示す。

また、**b_jは困難度パラメタ**で項目特性曲線の位置を表わし、パラメタ値が大きい方が、項目特性曲線が右寄りになり、困難度が高いことを表わす。

さらに、**a_jは識別力パラメタ**で項目特性曲線の立ち上がりの程度(勾配)を表わし、パラメタ値が大きい方が、項目特性曲線の立ち上がりが急になり、識別力が大きいことを表わす。

- この他に、識別力は全ての項目で等しく困難度のみをパラメタとする「1パラメタ・ロジスティック・モデル(ラッシュ・モデル)」、
- 多枝選択形式の項目における「あて推量」をパラメタ c_j として取り込んだ「3パラメタ・ロジスティック・モデル」などもよく用いられる。
- なお、ラッシュ・モデルは数理的には、このようにIRTモデルの中に位置づけることもできる。

実際のテストに項目応答理論を適用するには、

- まず項目パラメタの値を推定しておく必要があり、
受験者の測定結果は、受験者が各項目に対して正答したか誤答したかをまとめて**正誤パターン**で表わし、
それと解答した項目のパラメタ値とを用いて推定する。
- 具体的な手続きに関しては、例えば、野口・渡辺(2015)などを参照されたい。

IRTの特徴をまとめると・・・

- 1) 項目の困難度が受験者集団とは独立に特性尺度上の1点として定義される。
- 2) 受験者の特性(能力)尺度値が、解答した項目群とは独立に特性尺度上の1点として定義される。
- 3) 項目の困難度と受験者の特性(能力)尺度値とが同一の尺度上に位置付け表わされる。
- 4) 項目の特性は、特性尺度値と正答確率の関係を表わす「項目特性曲線」で全て記述される。

そして、

- 5) ある項目に正答するか誤答するかは、他の項目に正答したか誤答したかの影響を受けず、相互に独立である(局所独立の仮定)
- 6) 測定精度が特性(能力)尺度値の関数として表われ、尺度値ごとにきめの細かい測定精度の評価が可能になる。

などの特徴がある。

適用場面には・・・

- IRTはさまざまな場面での適用可能性があるが、特に
 テストの「等化(equating)」と
 テスト項目の「特異項目機能
 (Differential Item Functioning:DIF)」
を検出する場面で有用である。

項目応答理論 (IRT) と等化

- 複数のテストの測定結果を比較可能にする

= テストの等化

- より厳密には、

同一の能力特性を測定する目的を持ち、

同一の仕様 (specification) で開発されているが、

異なる問題項目から構成される複数のテストの得点を、

相互に比較可能な共通尺度上の得点に変換して表わす操作のこと

IRT尺度の等化

1) IRT尺度では、

項目パラメタ値や特性尺度値を表現する目盛りの原点と単位とは線形変換の範囲内で自由に決定できる。

2) 異なる原点と単位をもつ複数のIRT尺度は、

適切な線形変換を行うことによって、

すべて共通の原点と単位をもつ共通尺度に合わせられる。

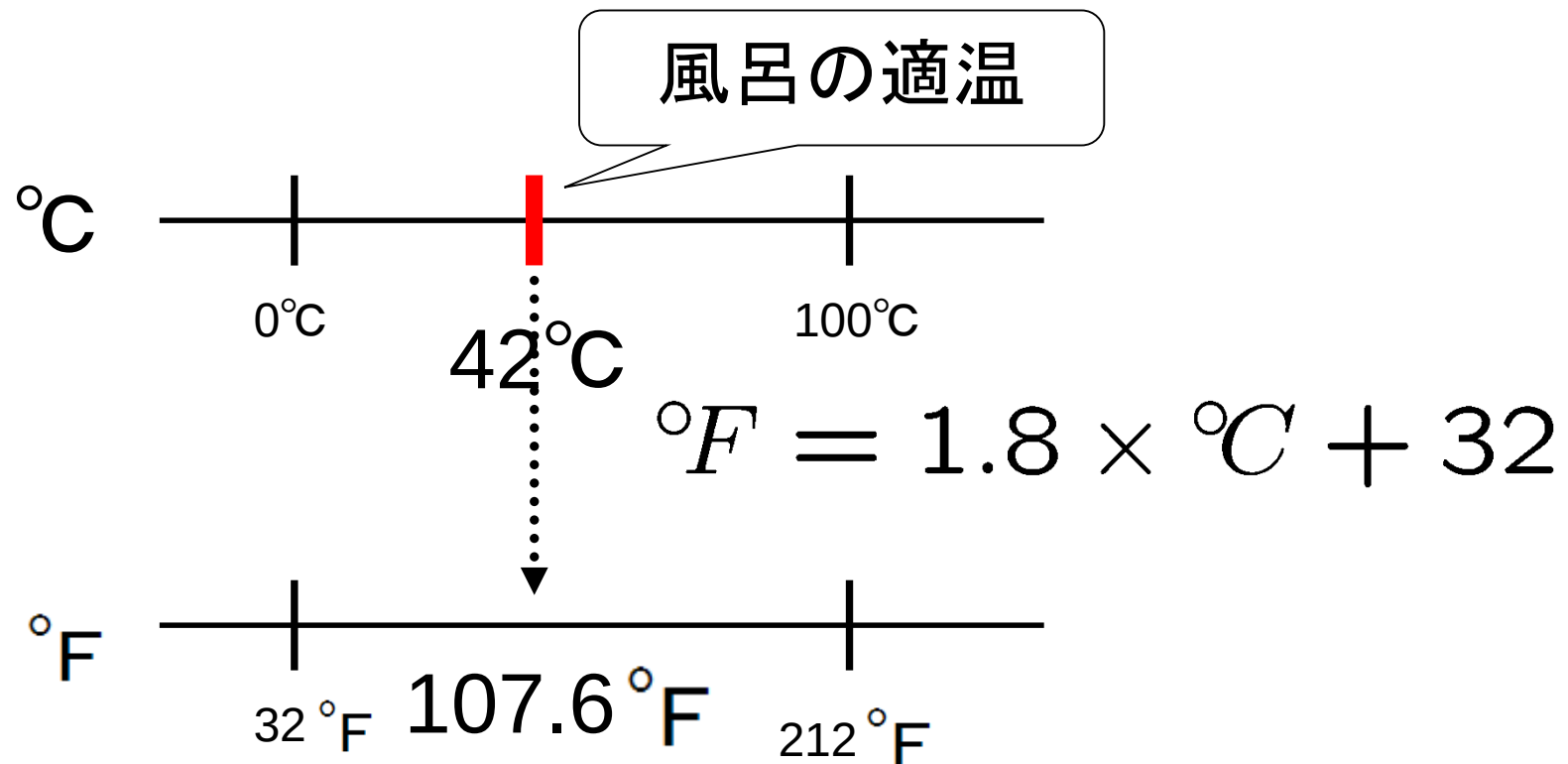
これを「IRT尺度の等化」という。

すなわち、IRT尺度 θ に

$$\theta^* = k \times \theta + l$$

という線形変換を行なっても、対応する尺度 θ^* 上で正答確率は不変であることが示される。 $(k \neq 0)$

このことは...



3. CBTとは何か

- Computerized test
コンピュータを利用したテスト
- 1 Computer Based Test ; CBT
 - 2 Computerized Adaptive Test ; CAT

Computer Based Test

- コンピュータを用いて実施するテスト方式
- コンピュータの可能性の拡大に対応して、テストの可能性を
広げる
- 項目 画像・音声・動画
 ある意味で真正性の向上
 実技試験にも用いられる
- テストの進行 応答時間の記録・項目の最適化

- 採点システム 結果の即時的返却(対個人)
集団データの分析(対機関)
- しかし、あくまでテストの可能性を拡げるもの
PBTにとってかわるものではない
必要に応じて冷静に使い分けることが必要
- 技術面での信頼性をきちんと評価することが重要
- テストは媒体も大切だが、測定する内容が最も大切なことを忘れてはいけない！ ➡ 測定の妥当性検証

Computerized Adaptive Test

- 受験者個人毎に最適な項目を選択・編集して実施するテスト方式
 - ➡ すべての受験者に対して一定水準の精度の測定が可能になる
- 基本的には、
直前に実施した項目に対して受験者が
正答した場合には、次にはより難しい項目を、
誤答した場合には、より易しい項目を
実施する、という手続きを繰り返す。

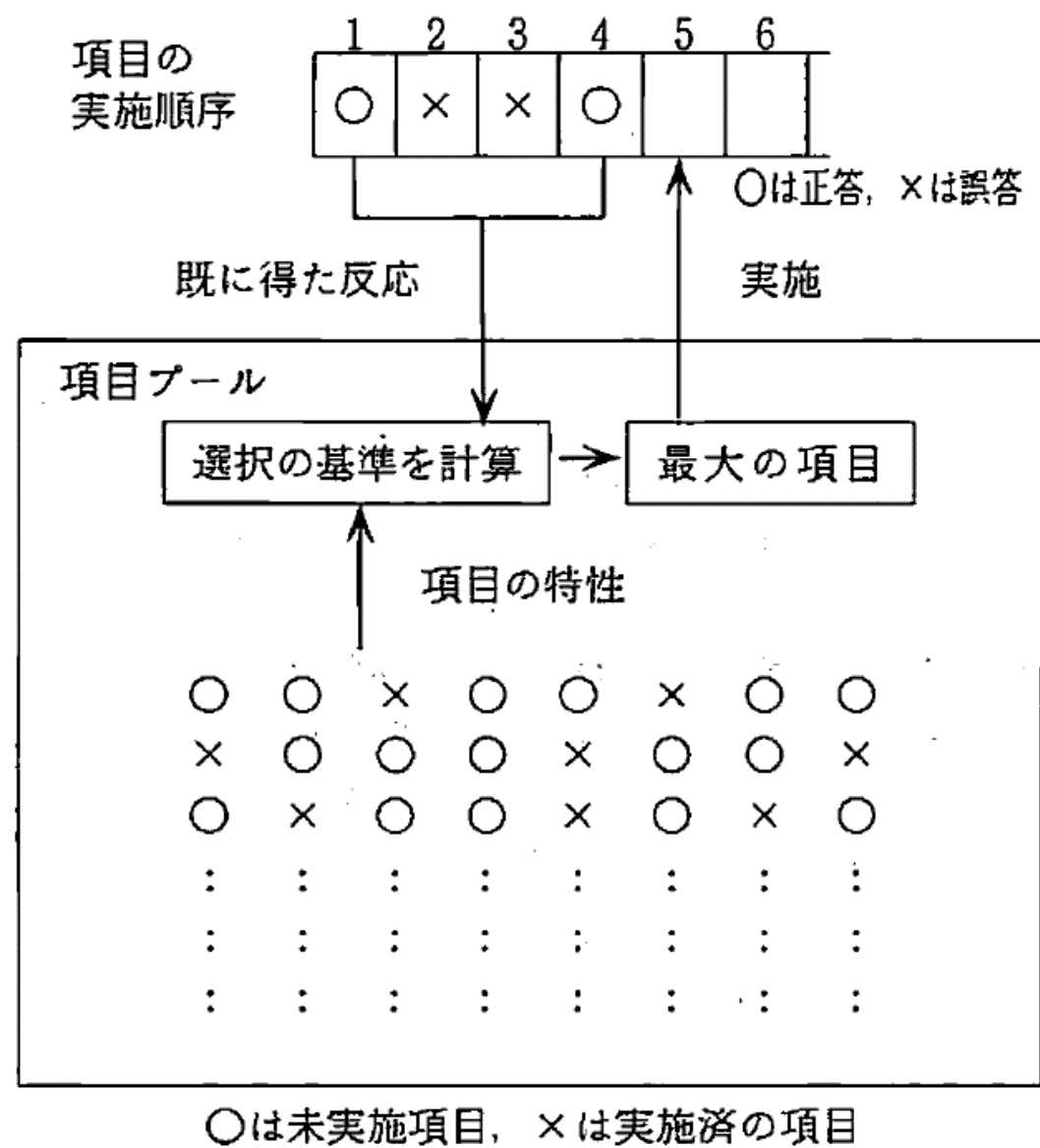
- CATでは、
受験者毎に解答する項目
解答項目数
が異なる。

それにもかかわらず、測定結果を同一の尺度上に表現して、
測定精度を確認できる。

➡IRTの適用が基本的条件

CATの利点

1. すべての受験者に対して高い精度の測定が実施可能
2. 精度を落とすことなく、受験者一人あたりに実施する項目数を減らすことができる
 - ➡ テスト実施時間を節約できる
3. 難しい項目が続いて、受験者にフラストレーションや不安を起こさせたり、易しい項目が続いて飽きさせたりすることがない
4. 受験者の予定に最大限配慮して実施できる
 - ➡ ただし、膨大な数の項目から構成される項目プールが必要



CBT化にあたって

- High stakes test の場合には、高い技術の信頼性が要求される。
- CBTを実施する施設の整備は十分か。
- IRTが必須ではないが、適用可能なテスト仕様になっているか。
- 膨大な数の項目が蓄えられた項目プールが整備可能か。
- テストの実施状況を管理する人材が確保できるのか。
- 特別な配慮を必要とする受験者に対応できるのか。
- 技術の進歩に対応して、テスト開発をする、人的資源、時間、資金は十分用意できるのか。

4. 英語4技能外部試験をどう位置づけるか

- 高大接続改革で最も話題に上るのが「大学入学共通テスト」。
- 当初はIRTとCBTが目立ったが、最近はトーン・ダウンしている。
- 「英語」の試験に関する改革は変わらず、大きな議論を巻き起こしている。
- 現在名前が挙がっている外部試験は、IRTベースかつCBTのものが多い。そういう意味では、IRTとCBTが活用されると言える。
- しかしながら、「共通テスト」と大学が独自に実施する「個別試験」との総合的な議論の影が薄くなっている点が気になるところである。

- 教科「外国語」の科目「英語」に関して、2020年度から2023年度までの4年間は「共通テスト」とTOEFLやCambridge英検などの大規模英語試験の中から大学入試センターが認定する「認定試験」とを併存させることになった。
- しかし、複数の認定された「英語試験」を各大学はどのように他の教科・科目と合わせて総合的に入学者選抜に用いればよいのかについては明らかにされていない。
 - ➡ 各大学で自由に考えていいのか？

CEFRによる「認定試験」の対応づけ？

- 大規模英語試験は、その目的も異なれば、テストの仕様も異なっている。
- どの試験が優れているというのではなく、「英語」のどういう能力を何のために測定するのかが異なっている。
- これに対して、それらのテストの測定結果をCEFRの能力レベルに対応づけた表が文部科学省から資料として公開されている。

CEFRとは？

- CEFRは、
Common European Framework of Reference for Languages
(欧州共通言語参照枠組み)を約したもの。
- 欧州域内の外国語教育専門家が、言語学習や教授法、
評価法に関して、国や言語の違いを超えて、相互理解
およびコミュニケーションを促進するための共通基盤となる
「言語参照枠組み」を提示した260頁に及ぶ文書で、2004年
に欧州評議会によって出版された。

- 従って、CEFRは異なるテスト間での測定結果を対応づけるものではなく、英語に限らず他の言語も共通に、学習者のその時点での言語能力を6段階のレベルで表わすもの。
従って、「フランス語」、「ドイツ語」、「中国語」などの大規模試験もひとつの表に入れることができる(表1参照)。
- なぜそのようなことができるのかと言うと、CEFRは個別言語で設定される「能力規準」ではなく、言語能力レベルのイメージを共有できるように、そのレベルの学習者がその言語を使ってどのようなことができるかを Can-do statements で表わした「参照枠組み」だからである。

- それにもかかわらず、大学入学共通テストでは英語以外の外国語科目をどのようにするのかについての議論が見えて来ない。

表 各種外国語能力試験とCEFRレベルの関係

CEFR	cambridge English	IELTS	TOEFL	英検	TOEIC/ TOEICS&W	GTEC CBT	DELF/ DALF	漢語水平考試
C2	CPE	8.5-9.0					DALF C2	筆記6級
C1	CAE	7.0-8.0	95-120	1級	1305-1390	1400-	DALF C1	筆記5級
B2	FCE	5.5-6.5	72-94	準1級	1095-1300	1250-1399	DELF B2	筆記4級
B1	PET	4.0-5.0	42-71	2級	790-1090	1000-1249	DELF B1	筆記3級
A2	KET	3.0		準2級	385-785	700-999	DELF A2	筆記2級
A1		2.0		3級-5級	200-380	-699	DELF A1	筆記1級

- また、各言語テストの測定結果をCEFRの6レベルに対応させることはできるが、言語テスト間の得点の対応関係を保証しているわけではない。
- そうなると、大学入学共通テストでは、認定試験の結果は、得点ではなくCEFRのレベルで表わすしかない。
- しかしながら、異なる英語のテストによる測定結果を比較可能な共通尺度上に表わす、「等化」という得点の変換はCEFRではできない。

高校3年生の英語能力

- 文部科学省による「英語教育改善のための英語力調査」の平成26年度調査の結果（全国から無作為抽出した国公立高校の3年生約70000名（「話す」技能は約20000名）の調査結果）。
- CEFRで「読むこと」が B2 0.2%、B1 2.0%、A2 25.1%、A1 72.7%、「聞くこと」がB2 0.0%(5名)、B1 0.7%、A2 12.8%、A1 86.5%であり、「話すこと」「書くこと」はそれよりも低い水準にとどまっていることが明らかになった。
 - ➡ 例えば、TOEFLではA2,A1レベルは測定の範囲外！

- 大学入学共通テストの受験者層と同質というわけではないが、大部分の受験生がA1、A2レベルにあることは確かであり、実力差が見えず、CEFRをそのまま導入しても入学試験で合否を決定する情報にはなり得ない。

5. 本日のまとめ

- 「大学入学共通テスト」に関して、IRTベースの試験の導入、CBTによる全国一斉実施という議論は現段階では実現せず、外部4技能英語試験の導入が最も現実味を帯びている。
- IRTベースの試験を開発するには、IRTが仮定する条件を満たす仕様をしっかりと組み立てることが必要。各教科、科目について、基礎的な研究を進める必要がある。
- CBTに関してはテクノロジーの進歩を大規模試験に取り入れるという姿勢・柔軟さは必要であるが、いきなり全国規模の共通テストに導入する前に、大学・学部・学科など小さな単位での実施を繰り返して、問題点を洗い出し、解決しておくことが必要。

- 「大学入学共通テスト」は英語以外の教科・科目も合わせた全体として統一的な理念、構成、仕様のもとで開発・実施されることが望ましい。英語で4技能を測定することが必要であるならば、民間の知恵や技術も導入しながら大学入試センターで新しい英語テストを開発することも考えていいのではないか。
- 「共通試験」と「個別入試」とを総合的に検討することが必要。共通一次試験導入の際には、一次試験と二次試験の役割のことがしっかり議論されていたように思う。

- 外部英語試験の導入は、「大学入学共通テスト」ではなく、各大学が教育理念や大学・学部・学科の特徴に応じて「個別試験」の中で適切なものを選択して利用することが望ましい。
- 高大接続改革の中で、大学入学試験をどうするかは大きな問題ではあるが、もはやすべての大学について一律に議論できる状況ではない。大学の個性に応じて利用しやすい「共通テスト」であっていいのではないか。
- 何よりも、テストの研究開発には、人的資源、時間、資金が必要であることを肝に銘じておかなければならない。